

採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル —— MCMC アルゴリズムに基づく推定 ——

宇佐美 慧*

小論文試験や面接試験、パフォーマンステストなどに基づく能力評価には、採点者ごとの評価点の甘さ辛さやその散らばりの程度、日間変動といった採点者側のバイアス、および受験者への期待効果、採点の順序効果、文字の美醜効果などの受験者側のバイアス要因の双方が影響することが知られている。本論文では Muraki (1992) の一般化部分採点モデルを応用して、能力評価データにおけるこれら 2 種類のバイアス要因の影響を同時に評価するための多値型項目反応モデルを提案した。また、母数の推定については、MCMC 法 (Markov Chain Monte Carlo method) に基づくアルゴリズムを利用し、その導出も行った。シミュレーション実験における母数の推定値の収束結果から推定方法の妥当性を確認し、さらに高校生が回答した実際の小論文評価データ (受験者 303 名, 採点者 4 名) を用いて、本論文で提案した多値型項目反応モデルの適用例を示した。

キーワード：項目反応理論, 能力評価, バイアス, 小論文試験, MCMC

問題と目的

能力評価における測定論上の問題

入学試験や入社試験をはじめとして、応用的な思考能力、表現力、実技能力、独創性などといった、受験者のもつ高次の能力を多面的に測定・評価することの必要性が近年特に強く指摘されている。そして客観式テストでは、これら高次の能力の測定・評価を行うことは多くの場合困難であることから、小論文試験、面接試験、パフォーマンステストを通した能力評価が行われることが一般的であり、その利用も広くなされている。例えば、大学入試の小論文試験については、客観式テストの過度な利用に伴う暗記主義・偏差値至上主義に歯止めをかける意図や (例えば大野木, 1994), また各大学における選抜方法・評価観点の多様化やテストのアカウンタビリティへの要請に対応する意図 (例えば平井, 2007) などから急速に広まり、現在は国公立大学および私立大学の 8 割が小論文試験を選抜に利用している (文部科学省, 2005)。

しかし、これらのテスト形式に基づく能力評価は、人間の多様な能力や心理的特性を評価するには一般に有効なテスト形式ではあるものの、一方でその評価の構成概念妥当性、信頼性、バイアス要因に関連する測定論上の問題が、特に小論文試験の場合を中心に広く

指摘されている (Barkaoui, 2007 ; Brown, Glasswell, & Harland, 2004 ; 平井, 2002 ; 平・江上, 1992 ; 宇佐美, 2008 ; 渡部・平・井上, 1988)。そして、これらの問題点については未だ十分に解明されていない点も多々あり、既に広く利用されている小論文試験の場合においても、その評価に伴う測定論上の問題については多くの点で検証が不十分であるのが現状である。(宇佐美, 2008 ; 渡部, 1994)。

能力評価データに伴う採点者側と受験者側のバイアス要因

能力評価データにおいて上記の測定論上の問題が伴う場合、その分析の際にもこれらの点を考慮したデータ収集デザインを採用することが望ましい。例えば、小論文試験、面接試験、パフォーマンステストのいずれの場合においても、評価の信頼性を維持する意図から複数の採点者により評価を行い、その結果評価データが受験者×項目の二元ではなく、受験者×項目×採点者の三元データとなる場合が多い。ところが、これは信頼性を改善するためには優れた収集デザインであるものの、例えば小論文評価の場合では、採点者間で評価の甘さ・辛さが存在することや、評価の際に力点を置く箇所が一貫しないこと、他にも評価点のバラつきが採点者間で異なることなどが指摘されており (e.g., 平井・渡部, 1994 ; 宇佐美, 2008 ; 渡部他, 1988), これら採点者側のバイアス要因を考慮した分析手法を採用することが望まれる。また類似したケースに、採点者は一名であっても、同様の意図から一枚の答案を複数回採点

* 東京大学大学院教育学研究科・日本学術振興会
〒113-0033 文京区本郷 7-3-1
usami_s@p.u-tokyo.ac.jp

する場合があります (e.g., Cronbach, Linn, Brennan, & Haertel, 1997; Penny, Johnson, & Gordon, 2000), この場合評価データは受験者×項目×採点時期の三元データになり, 同時に採点者内の気分の変化, 日間変動, 疲労などから生じる評価の一貫性の問題が生じ, これらも採点者側のバイアス要因になりうる。

また, 評価データに影響するバイアス要因は採点者側だけでなく受験者側に帰属されるものも幾らか知られており, それには, 例えば受験者の性別や期待効果, 評価の順序効果, さらに小論文試験の場合は文字の美醜効果などが挙げられる (e.g., Chase, 1983, 1986; Eames & Loewenthal, 1990; Hughes, Keeling, & Tuck, 1983; 宇佐美, 2008)。例えば Hughes et al. (1983) の実験においては, 単一の作文の評価において文字の美しい文章はそうでない文章に比べ, 採点の結果に 25 点満点中平均 1.32 点 (100 点満点であると約 5 点分) の差が生じたことが示されており, そのバイアスの影響の大きさがうかがえる。また, 多くの場合, 受験者側および採点者側に関するバイアス要因は同時にかつ複合的に影響することが多いと考えられることから, テストを通した能力評価の妥当性をより高めるには, これら 2 種類のバイアス要因の影響を共に考慮する必要があると言える。

項目反応理論との関連

項目プールを作成し当該のテストの受験者集団や項目の特性に直接依存しない能力評価を実現する目的や, また評価の信頼性を能力水準別に検証する目的で, さらには CAT (Computer Adaptive Test) の枠組みでテストを運用するために, 心理学・教育学・医学分野を中心に項目反応理論 (item response theory; IRT) の利用が進んでいる。Lawley (1943) や Lord (1952) らが IRT の理論的基礎を築き上げてから今日に至るまで, 項目反応を表現するモデルは多数提案されており, それらの具体的なモデルに関しては, 例えば Baker & Kim (2004), van der Linden & Hambleton (1997) 等に詳しい。

小論文試験, 面接試験, パフォーマンステストなどを通した能力評価データも, IRT の枠組みを用いた解析が可能である。例えば, 小論文評価データにおいては, 小論文などの論文体テスト項目が客観式テスト項目と共に出題される場合や, 他にも複数の評価観点に基づいて採点する分析的評価 (analytical evaluation) を複数の採点者が行う採点デザインでテストが実施された場合などにおいて, IRT の枠組みを用いた解析が利用できる。平井・渡部 (1994) でも, 評価データのカテ

ゴリ数と信頼性の関連を検討する目的で, IRT の段階反応モデル (Samejima, 1969) を用いた分析例が示されている。

他にも, 特に小論文評価研究の文脈では, 一般化可能性理論 (e.g., Brennan, 2001) が用いられることが多い (e.g., Braun, 1988; 梶井, 2001; 渡部他, 1988)。しかし, 一般化可能性理論では採点者や項目要因などから説明される分散成分の推定に力点が置かれているため, 例えば山内 (1999) でも指摘されているように, 採点者の個人差を明確にし, 問題となる特定の採点者や項目およびその組み合わせを検討するには不向きである。一方, IRT の場合, 個別の項目や採点者に関する母数の推定値を用いることによってこれらの点の評価が可能となる。

IRT において, 前小節において述べた, 能力評価データに伴う採点者側のバイアス要因を考慮する場合には, 例えば HLM (Hierarchical Linear Model; Raudenbush & Bryk, 2002) のモデリングを応用したマルチレベル項目反応モデル (e.g., Adams, Wilson, & Wu, 1997; Fox & Glas, 2001) や, 他にも多相ラッシュモデルをはじめとした, 採点者側のバイアスを直接考慮することのできる項目反応モデル (e.g., Fischer, 1983; Linacre, 1994; Patz & Junker, 1999) によって解析が可能である。マルチレベル項目反応モデルはテスト得点を受験者の能力から説明される分散と採点者からの分散に分けて評価することができ, 応用性の高いモデルである。しかし, このモデルは受験者母数における階層関係の構造化を基礎としているために, 採点者間で存在するバイアス要因を, 困難度や識別力などの各項目母数の観点から多面的に評価しにくいと, 特定の採点者に基づくテスト得点がどのような側面において問題を孕んでいるのか明確にすることはできず, 採点デザインを改良していく上では必ずしも十分ではない。また, Fischer (1983) や Patz & Junker (1999) のモデルや, Linacre (1994) の多相ラッシュモデルにおいても, これらは主に困難度母数に対する採点者のバイアスしか考慮されていない。さらに, これらの多くが二値データのための項目反応モデルであることを考慮すると, 小論文試験, 面接試験, パフォーマンステストのような能力評価データへの適用には困難を伴うことが予想される。

一方, 受験者側のバイアス要因の影響を IRT の文脈において考慮する場合については, 例えば受験者の潜在特性値を何らかの項目反応モデルを用いて推定した後に, 事後的にバイアス要因と考えられる, 受験者に関する共変量データを利用した回帰モデルを構築して

その影響を評価する方法が考えられる。しかし、この方法では回帰係数の希薄化が生じることが既に知られているため(e.g., Usami, 2008), 共変量データを直接組み込んだ項目反応モデル(e.g., Usami, 2008; Zwinderman, 1991)を利用することにより, 回帰係数の希薄化を抑えながらバイアス要因の影響を除去した潜在特性値を同時に推定することが望まれる。ところが, これらのモデルにおいては上述の採点者側のバイアス要因については考慮されていないため, その適用には依然問題が残る。そのため, 採点者の個人差を多角的に捉えること, および共変量を用いた場合にその希薄化の影響を抑えることを同時に実現することのできるモデルが必要となるが, このような採点者に関するバイアス要因と受験者に関するバイアス要因を同時に考慮した項目反応モデルの検討は未だなされていない。

本論文の目的

本論文では, 小論文試験, 面接試験, パフォーマンステストなどの, 測定論上の問題を伴うとされる能力評価データにおいて, 採点者側と受験者側に帰属されるバイアス要因の影響を評価するための項目反応モデルを提案する。特に, データである評定値自体は多値であることが一般的であること, および項目母数の多様性を考慮して, Muraki (1992) のGPCM (Generalized Partial Credit Model) に基づいた構築を行う。また, 推定法については項目反応モデルに限らず, 多くの統計モデルにおける利用が一般的になりつつあるMCMC (Markov Chain Monte Carlo : e.g., Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) 法を利用し, その推定手順についても導出を行う。MCMCは後述のようにベイズ法に基づく推定法であり, パラメータに関する事前の知識を推定に反映させることができ, 小論文評価のための項目反応モデルにおいてはほとんど応用されていなかった方法論である。そして, 提案したモデルおよび推定方法の妥当性を検証するためのシミュレーションを行い, さらに実際の小論文評価データに対して提案した多値型項目反応モデルを用いた分析例を紹介する。

モデル

定式化

以下では受験者数を I , 項目数を J , 採点者数を R と設定し, さらに評価点のカテゴリ数についてはいずれの項目の場合でも共通に K と設定する。GPCMでは, 受験者 i ($1, \dots, I$) が項目 j ($1, \dots, J$) においてカテゴリ k ($1, \dots, K$) を選択する確率を, 以下のように表現す

る。

$$P_{jk}(\theta_i) = \frac{\exp[\alpha_j(k(\theta_i - \delta_j) - \sum_{m=0}^k \tau_{jm})]}{\sum_{l=0}^{K-1} \exp[\alpha_j(l(\theta_i - \delta_j) - \sum_{m=0}^l \tau_{jm})]} \quad (1)$$

ここで, θ_i は受験者母数であり, α_j , δ_j , τ_{jm} は項目 j (のカテゴリ m)における識別力, 困難度, 閾値に関する項目母数である。 α_j が高い項目は異なる潜在特性値をもつ被験者を明確に識別することができること, また δ_j が高い項目は被験者が高い番号のカテゴリを選択しにくいことをそれぞれ意味する。 τ_{jm} は潜在特性値と各カテゴリの選択確率の関係を表現する項目反応カテゴリ特性曲線 (Item Response Category Characteristic Curve : IRCCC) において, 隣接するカテゴリのIRCCCが交差する位置を意味する母数であり, 選択されたカテゴリの度数分布に関する情報も有している。また, 母数の識別性のために $\sum_k \tau_{jk} = 0$, $\tau_{j0} = 0$ という制約が一般に課せられる。このようにGPCMは識別力, 困難度, 閾値の3つの側面から項目に関する情報を評価することが可能である。

本論文で提案する項目反応モデルでは, 採点者と受験者側のバイアス要因の影響を同時に考慮するために, 受験者 i が採点者 r ($1, \dots, r, \dots, R$) から項目 j において k と評価される確率を

$$P_{rjk}(\theta_i) = \frac{\exp[\alpha_{jr}(k(\theta_i - \delta_{jr}) - \sum_{m=0}^k \tau_{jmr})]}{\sum_{l=0}^{K-1} \exp[\alpha_{jr}(l(\theta_i - \delta_{jr}) - \sum_{m=0}^l \tau_{jmr})]} \quad (2)$$

のように設定する。ここで, 項目母数については採点者要因を考慮して,

$$\alpha_{jr} = \alpha_j \alpha_r \quad \delta_{jr} = \delta_j + \delta_r \quad \tau_{jmr} = \tau_{jm} \tau_r \quad (3)$$

で表され, また受験者母数については, 標準化された共変量についてのデータ z_{is} を用いて,

$$\theta_i = \sum_s \beta_s z_{is} + \theta_i^* \quad (4)$$

と再母数化して表現する。 α_{jr} , δ_{jr} , τ_{jmr} は項目 j (のカテゴリ m)における採点者 r の評価の識別力, 困難度, 閾値に関する項目母数である。(3)式においては, α_{jr} , δ_{jr} , τ_{jmr} が項目 j (のカテゴリ m)における全採点者についての平均的な識別力, 困難度, 閾値の大きさを意味する α_j , δ_j , τ_{jm} と, 採点者の個人差を意味する母数 α_r , δ_r , τ_r との乗法もしくは加法モデルであることを意味している。 α_r , δ_r は, それぞれ評価点を決定する際の識別力の高さと, 評価の厳しさの個人差を表す母数であると解釈できる。すなわち, α_r の値が高いほど採点者がより明確・一貫した基準のもとで受験者を評価していること, および δ_r が高いほど採点者が総じて低い評価点をつけやすいことを, それぞれ意味している。そして, τ_{jmr} は採点者 r の項目 j の評価に基づくIRCCCにおいて, 隣接するカテゴリのIRCCCが交差

する位置を意味する母数であることから、 τ_r の値が高いほど、IRCCCの交差する点を中心($\theta=0$)から一様に離れるため、その結果評価点の分散が小さくなることがわかる。すなわち、 τ_r は与えた評価点の平均的な散らばりの程度を意味する、乗法的に作用する母数と解釈できる。さらに、GPCMの性質から $\sum_k \tau_{jkr}=0$, $\tau_{j0r}=0$ という制約が伴う。

また、これら(3)による再母数化をしなくとも、異なる採点者に基づく評価項目群を互いに異なる項目とみなして、例えばGPCMを用いてこのデータにおける項目母数を求め、その後事後的に採点者母数を推定する方法も考えられる。しかし、この場合本来では推定すべき母数が $J+R$ (τ の場合は $J \times K+R$)個であるにもかかわらず、 $J \times R$ (τ の場合は $J \times K \times R$)個の母数を推定する必要があるため、母数全体の推定精度の観点から問題が生じることは明らかである。他にも、このように事後的に採点者母数を推定する方法は、特に α_r や τ_r など項目母数に対して乗法的に作用している採点者母数において、その推定値の標準誤差を評価することが難しいといった問題点を孕んでいる。この問題点は、後で述べるMCMCに基づく推定法の場合でも同様である。

また(4)式においては、 θ_i が標準偏回帰係数 β_s と標準化された共変量 z_{is} の線形和から説明される項 $\sum_s \beta_s z_{is}$ と、バイアス要因の影響とは独立な受験者の能力を意味する θ_i^* に分解されることを意味している。ここで S は共変量の数を意味する。この再母数化により、希薄化の影響を抑えた上で β_s を推定することが可能になる。

なお、本論文では、 z_{is} は期待効果や順序効果、文字の美醜効果などの受験者側のバイアス要因の影響を意味する共変量であるが、Zwinderman (1991)やUsami (2008)のように受験者の家庭環境や性別、学習状況などの独立変数と受験者母数との関連を検証する意図で、(4)式と同様の回帰式が構成される場合もあるため、これらの目的での応用も可能である。ただしこの場合、 θ_i^* はバイアス要因の影響を除いた受験者の能力という意味づけではなく、独立変数からでは説明されない残差の大きさを意味する値と解釈できる。

これらより、(2)式は採点者を表す添え字を除けば(1)式のGPCMと等価であるが、(3)と(4)の再母数化により、採点者側および受験者側のバイアス要因の影響を同時に評価できる点に本モデルの独自性があると言える。さらに、分析の目的に応じて、採点者母数に対して等値制約を課して推定すべき母数の数を節約するな

ど、分析者の持つ仮定を推定に反映させることも可能である。

最後に、母数の識別性のために

$$\prod_r \alpha_r = 1, \sum_r \delta_r = 0, \prod_r \tau_r = 1, \quad (5)$$

および

$$\sum_i \theta_i = 0 \quad \sum_i \theta_i^2 = 1, \\ \sum_i \theta_i^* = 0, \sum_i z_{is} \theta_i^* = 0 \quad (1, \dots, s, \dots, S) \quad (6)$$

が仮定される。ここで、上の式は θ_i についてはそれが標準化されていること、および S 個の共変量とは無相関であるという制約を意味する。

母数の推定

項目反応モデルの母数推定においては最尤法が用いられることが多いが、近年では、主に母数の事前知識を推定に反映させる目的や、サンプルサイズが少ない場合にもそれに直接影響されない妥当な母数推定を行うために (Goldstein, 2003; Hox, 2002), ベイズ法の利用が盛んになっている (e.g., Baker & Kim, 2004; Patz & Junker, 1999)。ベイズ法では母数に関する知識を事前分布 (prior distribution) の形で反映させ、さらに尤度関数と合わせて、各母数に関する事後分布 (posterior distribution) に基づいて推定値を得る。とりわけMCMCは、ベイズ統計において事後分布や事後確率を評価するためのシミュレーション法であり (大森, 2001), 項目反応理論の文脈でも既に広く使われている (e.g., Albert, 1992; de la Torre, Stark, & Chernyshenko, 2006; Johnson & Junker, 2003; Patz & Junker, 1999)。能力評価データの規模によっては十分なサンプルサイズが確保されない場合も大いに考えられることから、ベイズ法に基づく推定法は有用であると考えられる。本論文ではMCMC法の中でも代表的なMetropolis-Hastingsアルゴリズムに基づいてその母数推定を行い、以下でその方法を説明する。なおMCMCの理論的な側面については例えばGelman, Carlin, Stern, & Rubin (2003)や伊庭・種村・大森・和合・佐藤・高橋 (2005)を、応用的な側面については例えばde la Torre et al. (2006), 豊田 (2008)などを参照のこと。

事前分布

本論文では、推定する母数 $\alpha_j, \alpha_r, \delta_j, \delta_r, \tau_{jk}, \tau_r, \beta_s, \theta_i^*$ についてそれぞれ以下のように事前分布を設定する。

$$\begin{aligned} \ln \alpha_j &\sim N(\mu_a, \sigma_a^2), & \ln \alpha_r &\sim N(1, \sigma_a^2) \\ \delta_j &\sim N(\mu_\delta, \sigma_\delta^2), & \delta_r &\sim N(0, \sigma_\delta^2) \\ \tau_{jk} &\sim N(\mu_\tau, \sigma_\tau^2), & \tau_r &\sim N(1, \sigma_\tau^2) \\ \beta_s &\sim N(\mu_\beta, \sigma_\beta^2), & \theta_i^* &\sim N(0, \sigma_\theta^2) \end{aligned} \quad (7)$$

(7)式では、 α が対数正規分布に従い、他の母数が正規分

布に従うことを表している。また、分布の平均値があらかじめ設定されている母数は(5)・(6)式における制約と直接対応している。そして、 $\mu_\alpha, \mu_\delta, \mu_\tau, \mu_\beta, \sigma_{\alpha_j}^2, \sigma_{\alpha_r}^2, \sigma_{\delta_j}^2, \sigma_{\delta_r}^2, \sigma_{\tau_j}^2, \sigma_{\tau_r}^2, \sigma_\beta^2, \sigma_\theta^2$ はいずれも事前に設定する超母数(hyper-parameter)と呼ばれる値である。この場合、分散部分の値が大きいくほど、事前知識が少なく母数に対する情報が曖昧であることを意味する。また、これらの事前分布については、de la Torre et al. (2006) にもあるように、ベータ分布の範疇で扱うことも可能である。

尤度関数と完全条件付事後分布

D をサイズ $I \times (J \times R)$ の項目反応行列、また Z をサイズ $I \times S$ のバイアス要因の共変量に関するデータ行列とする。記法の簡便のため、 $\alpha_{j*} = (\alpha_1, \alpha_2, \dots, \alpha_j)'$, $\alpha_{r*} = (\alpha_1, \alpha_2, \dots, \alpha_r)'$, $\delta_{j*} = (\delta_1, \delta_2, \dots, \delta_j)'$, $\delta_{r*} = (\delta_1, \delta_2, \dots, \delta_r)'$, $\tau_{j*k} = (\tau_{1k}, \tau_{2k}, \dots, \tau_{jk})'$, $\tau_{j*} = (\tau'_{j*1}, \tau'_{j*2}, \dots, \tau'_{j*K})'$, $\tau_{R*} = (\tau_1, \tau_2, \dots, \tau_R)'$, $\alpha = (\alpha_{j*}', \alpha_{R*}')'$, $\delta = (\delta_{j*}', \delta_{R*}')'$, $\tau = (\tau_{j*}', \tau_{R*}')'$, $\beta = (\beta_1, \beta_2, \dots, \beta_S)'$, $\theta = (\theta_1^*, \theta_2^*, \dots, \theta_I^*)'$ (ここで'は行列の転置を意味する) と表記すると、データ D , Z が得られたときの尤度関数は、

$$L(\alpha, \delta, \tau, \beta, \theta | D, Z) = \prod_{i=1}^I \prod_{j=1}^J \prod_{r=1}^R P_{rjk}(\theta_i) \quad (8)$$

となる。そして、母数間の独立性を仮定すると、事後分布は、

$$P(\alpha, \delta, \tau, \beta, \theta | D, Z) \propto L(\alpha, \delta, \tau, \beta, \theta | D, Z) P(\alpha) P(\delta) P(\tau) P(\beta) P(\theta) \quad (9)$$

のように単に尤度関数と各事前分布を掛け合わせた値となる。ここで $\propto$ は比例を意味する記号である。また $P(\alpha), P(\delta), P(\tau), P(\beta), P(\theta)$ は(7)式に基づく事前分布であり、例えば $p(\alpha)$ の具体的な形は、

$$P(\alpha) = p(\alpha_{j*}) p(\alpha_{R*}) = \prod_{j=1}^J p(\alpha_j) \prod_{r=1}^R p(\alpha_r) \quad (10)$$

である。したがって各母数の、他の全ての母数に関する条件付確率を意味する完全条件付事後分布は、例えば α の場合は、

$$P(\alpha | \delta, \tau, \beta, \theta, D, Z) \propto L(\alpha, \delta, \tau, \beta, \theta | D, Z) P(\alpha) \quad (11)$$

のように、尤度と事前分布の積に比例した形で表現できる。ここまでの議論に関するより詳細な点については、例えば de la Torre et al. (2006) を参照のこと。

Metropolis-Hastings アルゴリズム

本小節では、Metropolis-Hastings アルゴリズムに基づいた推定手順を説明する。MCMC は反復計算を必要とするアルゴリズムであるため、初期値を何らかの方法で設定する必要がある。例えば、識別力母数の場合は当該項目における評価点と合計得点との相関係

数を利用する方法などがある。詳細については例えば Gelman et al. (2003) や de la Torre et al. (2006) など を参照のこと。

ここでも、まず α の場合に基づいて説明する。初期値を設定したら、 α^t を t 回目の α の推定値ベクトルとしたとき、更新の候補となるベクトル α^* を、 $N(\alpha^t, \sigma_{\alpha^*}^2 I_J)$ から抽出する。ただし、 I_J はサイズ J の単位行列であり、 $\sigma_{\alpha^*}^2$ は任意に設定される分散パラメータである。この候補値を抽出する際に用いる分布を提案分布と言い、ここでは上記のように多変量正規分布を用いているが、分散成分に単位行列を利用していることから、結果的に各母数は互いに独立に抽出されていることになる。そして、次の更新値である α^{t+1} は以下の確率で $\alpha^{t+1} = \alpha^*$ と更新される。

$$\min \left[\frac{P(\alpha^* | \delta, \tau, \beta, \theta, D, Z)}{P(\alpha^t | \delta, \tau, \beta, \theta, D, Z)}, 1 \right] \quad (12)$$

ここで $\min$ は小さい方の要素を選択することを意味する記号である。もし α^* が採択されなかった場合、 $\alpha^{t+1} = \alpha^t$ と値は変わらない状態で更新される。 $\delta, \tau, \beta, \theta$ も同様の方法で更新される。ただし、 $\alpha_{R*}, \delta_{R*}, \tau_{R*}$ は(5)の平均値の制約が、加えて θ は(6)の共変量との無相関性の制約があるため、若干の工夫が必要である。 $\alpha_{R*}, \delta_{R*}, \tau_{R*}$ の場合では、候補値を抽出したあと、制約を満たすように線形変換したものを候補値ベクトルとすればよい。 θ についてはさらに共変量データ行列の射影行列を用いて、共変量からは説明できない部分を抽出すればよい。すなわち、 θ^t を t 回目の推定値ベクトルとし、 θ^* を $N(\theta^t, \sigma_{\theta^*}^2 I_I)$ から抽出したときに、まず θ^* を平均 0 になるように変換し、次に

$$\theta^{**} = [I - Z(Z'Z)^{-1}Z']\theta^* \quad (13)$$

を計算する。そして(13)の操作による分散の変動を調整するために

$$\theta^{***} = \frac{\sqrt{\text{Var}(\theta^*)}}{\sqrt{\text{Var}(\theta^{**})}} \theta^{**} \quad (14)$$

と計算する。 $\text{Var}(\cdot)$ は分散を意味する。最後に、(7)式における $\sum_i \theta_i^2 = 1$ の制約を満たすために、

$$\theta^{****} = \frac{\theta^{***}}{\sqrt{\text{Var}(\sum_{s=1}^S \beta_s z_{is} + \theta^{****})}} \quad (15)$$

と計算する。この θ^{****} について(12)式と同様に完全条件付事後分布を利用した式を構成すればよい。なお最後の(15)の操作は β の抽出の際にも行う。

母数の点推定値については抽出されたサンプルの平均値で、標準誤差はサンプルの標準偏差から推定することができる (Gelman et al., 2003)。また、事後分布からのサンプルとしての精度を高めるため、最初の B 回

分のサンプル結果は除外して推定を行うことが多い。
この B 回の期間のことを Burn-in と言う。

シミュレーション

方法

本節では、前節で述べた推定手順の妥当性を確認するために、項目数や採点者数、受験者数等の条件を変化させながら、適当に設定した母数の真値を利用して人工データを発生させ、そのデータに対して前節で述べた方法を適用して母数の復元の精度を検討した。具体的には、以下の手順に基づいて行った。

STEP 1 θ_i および共変量データ $Z(S=1)$ を標準正規乱数を利用して I 個発生させる。

STEP 2 項目数 J 、採点者数 R のもとで任意に設定した母数の真値より、 j^* を $1 \leq j^* \leq J \times R$ としたときに、サイズ $I \times (J \times R)$ の項目反応確率行列 $P_{j^*k}(\theta_i)$ をカテゴリ数の $K=5$ 個分計算する。

STEP 3 項目反応確率行列と同じサイズの一様乱数行列 U を発生させ、以下の要領で項目反応データ D を作成する。

$$D_{ij^*} = \begin{cases} 5(U_{ij^*} > \sum_{k=1}^4 P_{j^*k}(\theta_i)) \\ 4(\sum_{k=1}^4 P_{j^*k} \geq U_{ij^*} > \sum_{k=1}^3 P_{j^*k}) \\ 3(\sum_{k=1}^3 P_{j^*k} \geq U_{ij^*} > \sum_{k=1}^2 P_{j^*k}) \\ 2(\sum_{k=1}^2 P_{j^*k} \geq U_{ij^*} > P_{j^*1}(\theta_i)) \\ 1(P_{j^*1}(\theta_i) \geq U_{ij^*}) \end{cases} \quad (16)$$

ここで各母数に関する真値 $\alpha_{jt}, \alpha_{rt}, \delta_{jt}, \delta_{rt}, \tau_{j1t}, \tau_{j2t}, \tau_{j3t}, \tau_{j4t}, \tau_{rt}, \beta$ はそれぞれ $\alpha_{jt} \sim U(0.4, 1.0), \alpha_{rt} \sim U(0.7, 1.4), \delta_{jt} \sim U(-1.5, 1.5), \delta_{rt} \sim U(-0.7, 0.7), \tau_{j1t} \sim N(-1.2, 0.5), \tau_{j2t} \sim U(-0.8, 0.5), \tau_{j3t} \sim U(0.8, 0.5), \tau_{j4t} \sim U(1.2, 0.5), \tau_{rt} \sim U(0.7, 1.4), \beta \sim U(0.2, 0.5)$ から互いに独立に発生させた。ここで U は一様分布を意味する。採点者母数の定義域からもわかるように、極端ではないものの、採点者間の評価の個人差が存在するものとして仮定されている。

また、STEP 2 の段階で設定されているように、本節のシミュレーションでは、採点は全ての採点者が全ての受験者および項目について評価する完全クロスデザインであること、および評価の局所独立性が成り立つことが仮定されている。シミュレーションでは、 $I=100, 300, 500, 1000$, $J=2, 5, 10$, $R=2, 4, 6$ の組み合わせの、計 36 ($=4 \times 3 \times 3$) 条件を実施した。事前分布については、 $\ln \alpha_j \sim N(0.7, 0.5)$, $\ln \alpha_r \sim N(1, 0.5)$, $\delta_j \sim N(0,$

$2)$, $\delta_r \sim N(0, 1)$, $\tau_{jk} \sim N(0, 2)$, $\tau_r \sim N(1, 0.8)$, $\beta_s \sim N(0.1, 0.3)$, $\theta_i^* \sim N(0, 0.99)$ と設定した。これらの事前分布は、分散を高め設定しているため、項目母数や採点者母数、回帰係数について事前にほとんど情報を持っていない状況であることを意味する。

そして、burn-in 期間については $B=20000$ 回を定め、反復は計 50000 回行った。収束の判断については様々な方法が知られているが (Geweke, 1992; Gelman et al., 2003; 大森, 2001), 本論文では一つの長めの連鎖を利用した収束判断の方法、および幾らか初期値を変化させた上での複数の連鎖に基づく推定値の収束度合いから検証する方法を併用し、上記の諸条件の場合において、サンプルの変動が十分に安定していることを確認した。

結果

推定方法の妥当性の検証方法はいくつか考えられるが、ここでは、各 36 条件において、それぞれ適当に設定された真値のもとで、50 回シミュレーションを繰り返し、各母数のバイアスおよび平均平方誤差 (Root Mean Squared Error; $RMSE$) を計算する。例えば、 α_j の場合はバイアスと $RMSE$ はそれぞれ以下ようになる。

$$\alpha_{j\text{bias}} = \frac{1}{50} \sum_{c=1}^{50} \alpha_{jc} - \alpha_{jt} \quad \alpha_{j\text{rmse}} = \sqrt{\frac{1}{50} \sum_{c=1}^{50} (\alpha_{jc} - \alpha_{jt})^2} \quad (17)$$

α_{jc} は c (≤ 50) 回目のシミュレーションにおける α_j の推定値である。

なお紙幅の都合から、ここでは全ての母数に関するバイアスと $RMSE$ を報告するのではなく、 $\alpha_{j*}, \alpha_{R*}, \delta_{j*}, \delta_{R*}, \tau_{j*}, \tau_{R*}, \beta$ の母数群ごとにそれぞれバイアスの絶対値に関する平均値、および $RMSE$ の平均値を計算した結果を Table 1 に報告する。ここで、例えば α_{j*} の場合なら以下ようになる。

$$\alpha_{j* \text{bias}} = \frac{1}{J} \sum_{j=1}^J |\alpha_{j\text{bias}}| \quad \alpha_{j* \text{rmse}} = \frac{1}{J} \sum_{j=1}^J \alpha_{j\text{rmse}} \quad (18)$$

GPCM やその拡張モデルにおいては、 α, δ の推定は比較的安定している一方で、 τ の $RMSE$ が大きくなりやすいことが知られているが (e.g., de la Torre et al., 2006; Usami, 2008), 項目母数においては今回のシミュレーションの結果でも、 $RMSE$ の値から示唆されるように、同様の傾向が見出されている。他の項目反応モデルの場合における一般的傾向と同様に、今回設定した程度の条件であれば、項目数や採点者数が多いほど、そして特に受験者数が多いほど安定した推定が実現しており、バイアスの値や $RMSE$ の値が小さくなっていることがわかる。全体的に、推定値は真値に近い値

Table 1 各母数群内の各要素に関して推定されたバイアスの絶対値の平均値 (RMSEの平均値)

採点者数	項目数	受験者数	α_{j*}	δ_{j*}	τ_{j*}	α_{R*}	δ_{R*}	τ_{R*}	β
R=2	J=2	I=100	.0563(.1734)	.0439(.1014)	.1409(.2965)	.0496(.1098)	.0376(.0877)	.0320(.0611)	.0949(.1121)
		I=300	.0537(.1165)	.0187(.0652)	.0624(.1492)	.0184(.0822)	.0297(.0439)	.0226(.0527)	.0850(.0890)
		I=500	.0539(.1075)	.0315(.0568)	.0591(.1276)	.0175(.0805)	.0152(.0353)	.0291(.0439)	.0753(.0806)
		I=1000	.0484(.0776)	.0214(.0347)	.0381(.0843)	.0117(.0505)	.0171(.0184)	.0238(.0354)	.0698(.0811)
	J=5	I=100	.0178(.1565)	.0311(.1372)	.0718(.2521)	.0186(.0734)	.0153(.0510)	.0056(.0642)	.0512(.0521)
		I=300	.0208(.0717)	.0091(.0651)	.0412(.1521)	.0129(.0416)	.0153(.0268)	.0085(.0561)	.0496(.0531)
		I=500	.0160(.0554)	.0210(.0742)	.0307(.1301)	.0030(.0259)	.0122(.0206)	.0051(.0259)	.0351(.0499)
		I=1000	.0116(.0337)	.0059(.0267)	.0221(.0683)	.0089(.0237)	.0032(.0154)	.0021(.0208)	.0389(.0298)
	J=10	I=100	.0348(.1205)	.0300(.0937)	.0471(.2618)	.0061(.0508)	.0071(.0328)	.0083(.0465)	.0349(.0515)
		I=300	.0116(.0673)	.0149(.0634)	.0319(.1440)	.0108(.0393)	.0049(.0180)	.0045(.0249)	.0309(.0343)
		I=500	.0149(.0664)	.0216(.0634)	.0362(.0939)	.0024(.0242)	.0057(.0114)	.0036(.0199)	.0338(.0359)
		I=1000	.0129(.0581)	.0049(.0479)	.0216(.0839)	.0090(.0165)	.0004(.0077)	.0014(.0149)	.0347(.0354)
	J=2	I=100	.0261(.0855)	.0293(.0948)	.0568(.1476)	.0436(.1047)	.0121(.0766)	.0265(.0888)	.0439(.0539)
		I=300	.0252(.0587)	.0230(.0758)	.0456(.1098)	.0236(.0901)	.0049(.0744)	.0109(.0841)	.0477(.0522)
		I=500	.0336(.0527)	.0153(.0349)	.0288(.0629)	.0159(.0804)	.0028(.0334)	.0067(.0429)	.0562(.0596)
		I=1000	.0241(.0520)	.0119(.0216)	.0244(.0481)	.0070(.0461)	.0029(.0226)	.0136(.0313)	.0378(.0392)
R=4	J=5	I=100	.0243(.0804)	.0270(.0894)	.0507(.1558)	.0225(.0731)	.0163(.0483)	.0168(.0718)	.0336(.0457)
		I=300	.0178(.0486)	.0094(.0438)	.0236(.0956)	.0054(.0381)	.0087(.0317)	.0053(.0551)	.0307(.0342)
		I=500	.0219(.0393)	.0139(.0423)	.0259(.0815)	.0067(.0325)	.0087(.0265)	.0054(.0311)	.0356(.0383)
		I=1000	.0085(.0262)	.0068(.0337)	.0240(.0654)	.0059(.0202)	.0018(.0246)	.0064(.0225)	.0283(.0319)
	J=10	I=100	.0232(.0779)	.0209(.0695)	.0304(.1314)	.0076(.0535)	.0038(.0408)	.0152(.0375)	.0337(.0391)
		I=300	.0141(.0542)	.0192(.0726)	.0339(.1177)	.0050(.0324)	.0068(.0223)	.0042(.0340)	.0263(.0289)
		I=500	.0192(.0577)	.0248(.0498)	.0224(.0797)	.0083(.0265)	.0025(.0153)	.0026(.0261)	.0402(.0406)
		I=1000	.0192(.0562)	.0072(.0377)	.0240(.0455)	.0089(.0212)	.0024(.0046)	.0054(.0225)	.0154(.0346)
	J=2	I=100	.0160(.0636)	.0348(.0828)	.0317(.1139)	.0322(.0939)	.0244(.0777)	.0248(.0638)	.0413(.0430)
		I=300	.0131(.0416)	.0074(.0429)	.0103(.0767)	.0136(.0755)	.0095(.0515)	.0125(.0798)	.0305(.0365)
		I=500	.0204(.0495)	.0177(.0321)	.0416(.0691)	.0037(.0606)	.0149(.0495)	.0079(.0601)	.0369(.0308)
		I=1000	.0101(.0339)	.0175(.0236)	.0099(.0298)	.0096(.0424)	.0034(.0246)	.0126(.0436)	.0242(.0158)
	J=5	I=100	.0173(.0533)	.0127(.0681)	.0202(.1051)	.0201(.0627)	.0068(.0512)	.0097(.0654)	.0314(.0357)
		I=300	.0151(.0423)	.0151(.0422)	.0135(.0831)	.0139(.0523)	.0084(.0373)	.0132(.0543)	.0226(.0269)
		I=500	.0148(.0307)	.0174(.0412)	.0132(.0583)	.0047(.0351)	.0049(.0255)	.0069(.0365)	.0239(.0260)
		I=1000	.0089(.0134)	.0063(.0226)	.0184(.0635)	.0137(.0199)	.0032(.0186)	.0059(.0331)	.0127(.0181)
	J=10	I=100	.0156(.0543)	.0234(.0449)	.0297(.0739)	.0078(.0561)	.0090(.0351)	.0099(.0395)	.0273(.0298)
		I=300	.0126(.0394)	.0112(.0369)	.0064(.0720)	.0039(.0318)	.0063(.0166)	.0051(.0256)	.0137(.0163)
		I=500	.0136(.0276)	.0213(.0318)	.0203(.0513)	.0064(.0302)	.0046(.0109)	.0051(.0201)	.0143(.0254)
		I=1000	.0085(.0134)	.0107(.0288)	.0118(.0420)	.0055(.0170)	.0055(.0072)	.0041(.0161)	.0081(.0103)

* $\alpha_{j*}=(\alpha_1, \dots, \alpha_j)'$, $\alpha_{R*}=(\alpha_1, \dots, \alpha_R)'$, $\delta_{j*}=(\delta_1, \dots, \delta_j)'$, $\delta_{R*}=(\delta_1, \dots, \delta_R)'$, $\tau_{j*}=(\tau_1, \dots, \tau_j)'$, $\tau_{R*}=(\tau_1, \dots, \tau_R)'$ であり, また β は回帰係数を意味する。

が得られており, 前節の MCMC に基づく推定方法が概ね満足できる結果を示しているといえよう。また, 採点者の個人差をより大きくした場合についても検討してみたが, 推定する母数の数そのものは基本的に変わらないため, ここで示した結果に対して一貫して異なった傾向は見出されなかった。

分 析 例

本節では, 実際の小論文評価データを用いた分析例を示す。

データについて

受験者: 秋田県の県立高校の2年生 303 名 (男子 155 名, 女子 148 名) 8 クラス。

課題: 課題A・課題Bの2つの小論文課題を実施した。課題Aは「小学校の授業における英語の早期教育は必

要であるか否かについて, あなたの意見とその根拠が明確になるように論述しなさい。」という, 小論文のテーマのみを与えられる形式であり, 課題Bは, 日本の親の子育ての態度に関する3つのデータをみて, 「日本の親の子育ての態度に関して, どのような客観的特徴がデータから読み取れるか。その内容を要約し, また望ましい子育てのあり方についてのあなたの考えを, 合わせて論述しなさい。」という, データの要約と意見の論述を求める形式である。他の研究 (宇佐美, 2009) の都合から, 303 名のうち 155 名(4クラス)は課題Aを 400 字, 課題Bを 800 字で, 残りの 148 名(4クラス)は課題Aを 800 字, 課題Bを 400 字で回答させている。ここでは制限字数の群分けを検証することが目的ではなく, また評価データの因子構造については群間での系統的な違いはないことが既に確認されているため, 本節で

は群分けの区別をせず 303 名のデータを分析する。

採点者：日常的に小論文の作成や評価を行っている専門家 2 名 (A・B) と高校国語教師 2 名 (C・D) の計 4 名。

評価観点の設定：Remondino (1959) や渡部他 (1988) を参考にし、小論文試験の作成や評価の専門家と協議をしながら、分析的評価の為の評価観点を 11 項目 (B 課題では 12 項目) 作成した。他の研究 (宇佐美, 2009) で評価構造に関する因子分析を行う目的のため、一般に比べ評価観点を多めに作成してある。しかし本節では提案した項目反応モデルに基づく分析例を示すことが目的であるために、ここでは「年齢相応の語彙力があり、表現が稚拙でないか。」に関する「語彙力」、「展開されている主張が説得的であり、納得できるか。」に関する「説得力」、「原稿用紙の正しい使い方・段落の設定・回答字数について問題はなかったか。」に関する「形式」、そして課題 B についてはこれら 3 つの他に、「要約が簡潔であり、グラフの内容が正確に読み取れていたか。」に関する「要約」の評価データを分析対象とした。したがって 1 名の採点者あたり 7 項目分のデータが含まれることになる。以後、課題 A における「語彙力」「説得力」「形式」の評価点を項目 1-3 として、課題 B については同様の順番で項目 4-6 として、そして「要約」は項目 7 として扱う。採点は、全ての分析的評価項目において 5 点満点で行った。

バイアス要因：冒頭で述べたように、文字の美醜効果は評価に不公平性を生じうるバイアス要因である。そこで、大学院生 2 名が文字の美しさを五段階で評定 (得点が高いほど小論文中の文字が美しいことを意味する) し、その平均値 ($M=2.97$, $SD=0.94$) を各受験者のバイアス要因に関する共変量データとした。また、回答字数が制限字数の半分に満たなかった 5 枚の文章の評価データについては除外して分析した。したがって $I=298$ とな

る。

記述統計

採点者別に、各 7 項目の評価点の平均値と標準偏差を以下の Table 2 に示す。まず項目について検討すると、全体に課題 A・B の「説得力」に対応する項目 2 および項目 5 の評価の平均値が低く、また「形式」に対応する項目 3 および項目 6 については逆に評価の平均値が高い。小論文の内容的な質と最も関連が深いと考えられる「説得力」については、受験者がまだ日常的に十分な小論文指導を受けていなかったことが大きな要因であると考えられる。「形式」については、原稿用紙の正しい使い方、段落の設定、回答字数などの基本的なポイントに誤りがなければ満点になる評価観点であったことが要因として考えられる。

評価者については、項目間でみたときの評価点の差の傾向は類似しているが、その平均値そのものには幾らかの違いがみられ、採点者による甘さ・辛さの違いが生じていることが考えられる。特に専門家である採点者 A・B については採点者 C・D に比べて評価点の平均値が低く、採点が相対的に辛いことがわかる。

また、全ての採点者と項目についての合計点に関するヒストグラムは Figure 1 のようになった。回答字数が半分以下であり不適切であるとみなされた答案は除外してあるものの、左に裾を引いた歪んだ得点分布になっていることがわかる。また、さらに 80 点以下の答案は、回答字数の点からは問題がないものの、採点者からみて、論の体を成しておらず受験者が真摯に回答したとは考えにくいとされる答案が大半であった。そこで、これらは実際の評価場面における採点データの例にはふさわしくないと判断し、以下の分析からは除外した。最終的に $I=289$ となった。

分析結果

事前分布については、前節のシミュレーションの場

Table 2 各採点者に関する各項目についての評定の平均値 (標準偏差)

	項目 1	項目 2	項目 3	項目 4	項目 5	項目 6	項目 7	全体
採点者 A	3.70 (0.77)	2.28 (1.05)	4.54 (1.04)	3.66 (0.77)	2.06 (1.01)	4.35 (1.13)	3.44 (1.03)	24.04 (3.65)
採点者 B	3.59 (0.72)	2.39 (1.00)	4.54 (0.94)	3.56 (0.77)	2.30 (1.13)	4.41 (0.99)	4.33 (1.02)	25.11 (3.36)
採点者 C	4.35 (0.87)	4.17 (1.09)	4.29 (1.27)	4.56 (0.75)	4.24 (0.91)	4.17 (1.20)	4.54 (0.85)	30.32 (4.14)
採点者 D	3.91 (0.52)	3.50 (0.70)	4.64 (0.72)	3.82 (0.52)	3.10 (0.61)	4.57 (0.79)	3.91 (0.76)	27.46 (2.68)
全体	15.54 (2.07)	12.34 (2.53)	18.01 (3.60)	15.60 (1.90)	11.71 (2.64)	17.50 (3.73)	16.22 (2.79)	106.92 (11.91)

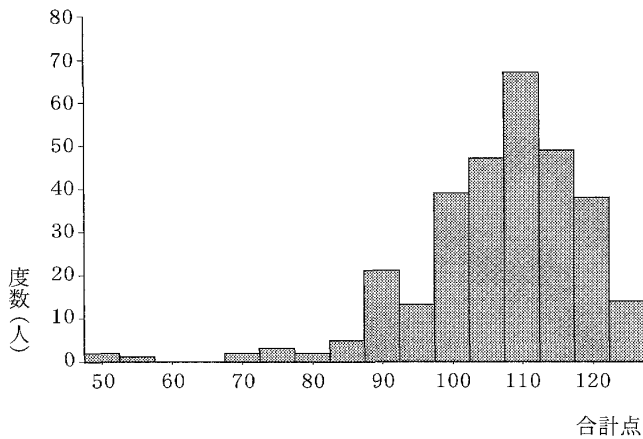


Figure 1 合計点のヒストグラム

合と同様の値を設定した。Burn-in および反復数についても同様に定めて、一つの長い連鎖を利用してサンプルの変動が安定していることを確認した。また、異なる4名の採点者に基づく各7項目分の評価データをそれぞれ異なる項目とみなした、計28項目からなるデータセットに対して因子分析（主因子法・プロマックス解）を適用した結果、第3固有値までの値は順に12.461, 3.752, 3.124（説明率は順に44.50%, 13.40%, 11.15%）であったため、概ね一因子性が保証されていると言える。

得られた項目反応データから推定された各母数の値は以下のTable 3のとおりである。また、紙幅の関係から τ については $\tau_{ji}(1, \dots, j, \dots, J)$ と τ_{R*} の推定値のみを記してある。

まず項目母数について検討する。 α_{J*} については、その平均値が高く天井効果を示していると考えられた「形式」にあたる、 α_3 と α_6 の値が相対的に小さい値に留まっている。これは、天井効果が生じた結果、個人の能力差を適切に判別ができなかったためであると考えられる。 δ_{J*} については、記述統計量の段階で示唆されていたように、「形式」の値（ δ_3, δ_6 ）は小さく、逆に「説得力」の値（ δ_2, δ_5 ）は高い値に推定されている。「語句」（ δ_1, δ_4 ）についても「形式」と同様に関わり低い値として推定されており、これらについては全体的に満点の評価が多く、天井効果の影響が出ていると考えられる。そして「要約」（ δ_7 ）についてはちょうど「形式」と「説得力」の中間程度の値が推定されている。今回用いた11種類（B課題は12種類）の分析的評価項目においては、全体的に評価点の平均値が高い項目が多かった中で、「説得力」は比較的全てのカテゴリにおいて得点が分布しており、項目と得点との相関が最も高い項目で

Table 3 項目母数、採点者母数、回帰係数の推定値および標準誤差

		推定値	SE			推定値	SE
α_{J*}	α_1	0.606	0.042	δ_{R*}	δ_7	-0.354	0.136
	α_2	0.703	0.041		δ_A	1.238	0.054
	α_3	0.351	0.027		δ_B	1.122	0.057
	α_4	0.714	0.043		δ_C	-1.801	0.010
	α_5	0.852	0.044	τ_{j1}	δ_D	-0.558	0.072
	α_6	0.275	0.026		τ_{11}	-4.326	0.263
	α_7	0.795	0.036		τ_{21}	-2.078	0.123
α_{R*}	α_A	1.411	0.047		τ_{31}	-1.319	0.324
	α_B	1.119	0.037	τ_{R*}	τ_{41}	-3.982	0.221
	α_C	0.603	0.028		τ_{51}	-1.795	0.102
	α_D	1.053	0.037		τ_{61}	-0.832	0.308
					τ_{71}	-1.011	0.257
δ_{J*}	δ_1	-2.693	0.122		τ_A	0.875	0.034
	δ_2	-0.291	0.056		τ_B	1.104	0.044
	δ_3	-3.481	0.201		τ_C	0.564	0.046
	δ_4	-2.562	0.108		τ_D	1.846	0.078
	δ_5	0.037	0.053		β	0.387	0.024
	δ_6	-3.458	0.220				

あった。 τ_{j1} については、特に「語句」の値（ τ_{11}, τ_{41} ）が低い値に推定されている。これは、「語句」の評価点の分布がほとんど3点、4点、5点の間に位置しているためである。一方、「語句」と同様に困難度の低い「形式」（ τ_{31}, τ_{61} ）においてはその得点分布が「語句」に比べて1,2点も含めた相対的に広い分布であったことから、「語句」よりも相対的に高い値に推定されている。

次に採点者母数について検討する。 α_{R*} に関しては全体的に専門家の採点者の方（ α_A, α_B ）が高く、これは専門家の採点者の方が良い文章と悪い文章を的確に識別して評価していたことを示唆するものであり、同時に評価の専門性の重要性をうかがうこともできる。 δ_{R*} については、記述統計の部分で示唆されていたように、特に専門家の採点者（ δ_A, δ_B ）において高い値に推定されており、評価が平均的により厳しくなっていることがわかる。 τ_{R*} については τ_C が低く、 τ_D が高めに推定されている。これは、採点者Dが特に3点、4点、5点の評価点が全体的に多く、評価点の分散が小さかったためであると考えられる。実際、 τ_{R*} の大きさは、前小節で検討した記述統計量における各採点者に関する合計得点の標準偏差の大きさと対応していることは興味深い。また、 τ_{R*} については採点者の専門性の違いによる明確な差が見られないこともわかる。しかし、今回 α_{R*} や δ_{R*} にみられた採点者の専門性の違いによる差については、採点者数がそれぞれ2名ずつであったという条件を考慮すれば、過度な一般化は慎むべきであろう。

今回、評価項目の選定やその操作化は採点者と共同で行い、また採点途中で評価基準が不明な点が生じた

場合には適宜採点者間で協議するなど、評価項目の反映する能力および各評価点の基準について共通の理解が得られるよう工夫をした。しかし、それでも冒頭で述べた多くの先行研究が示唆しているように、異なる採点者に基づく評価データには、程度の差こそはあれ採点者の個人差が反映されてしまう。しかも、従来の項目反応モデルで検討することのできた、評価点の平均値に関わる困難度母数だけではなく、識別力や閾値などにも個人差が強く反映されていた。特に、識別力母数については採点者間の専門性の差が反映されている可能性が高く、今回とは異なる受験者集団で同様の項目を用いる場合や、またより一般に、課題内容や目的とする能力特性が大きく変化しない範囲で異なる課題を実施する場合には、今回識別力母数の高かった採点者を優先して含めるなど、評価の精度が一定の基準に達するよう採点デザインを決定する必要があるだろう。また、困難度母数や閾値母数も含め、採点者母数は各採点者の与えた評価点において測定論的な観点から不適当と考えられる側面をある程度明確にすることができるため、評価点の決定の方法や評価手順に関する指導など、評価法の訓練を行う上でも有用であると思われる。

そして、文字の美醜効果の回帰係数については0.387と統計的に有意に高い値が推定されており、文字が美しい文章であるほど高い潜在能力値が推定されていることがわかる。ただ、文字の美醜効果と評価点の関係において単なる相関関係を超えた因果関係があるか否かを評価することは難しく、これらの点についてはまだ十分な議論がされているとは言えない。文字の美醜効果については、宇佐美(2008)でも指摘されているように、採点者や課題内容、他にも受験者の属性など多くの要因が複合的に影響していると考えられる。そのため、より因果関係の検証という観点から接近するには、これらの共変量を含めた上での評価が望ましいと言えよう。いずれにせよ、希薄化の影響を抑えた上で受験者に関するバイアス要因の影響の有無を見積もる目的においては、本研究で提案したモデルは有用であると言える。

総 括

本論文では、小論文試験、面接試験、パフォーマンステストなどの、測定論上の問題を伴うとされる能力評価データにおいて、採点者側と受験者側にあるバイアス要因の影響を同時に評価するための多値型項目反応モデルを提案した。従来、採点者間での評価データ

の特性に違いがあると思われた場合には、全採点者における平均データを用いたり、データを標準化したり、採点者ごとに個別に分析するなどの簡便法が取られることが多かったが、これはデータの持つ情報量を大きく損なうばかりか、推定そのものが不安定になるという欠点があったと言える。また、既存のモデルの枠組みで採点者母数を推定する場合には、例えばマルチレベル項目反応モデルや多相ラッシュモデルでは母数の種類の制約から採点者の持つ個人差を検討する上で必ずしも十分ではなかったこと、またGPCMを用いて事後的に採点者母数を推定する方法においてもその推定精度や、標準誤差を得る段階で問題点があった。

本論文で提案されたモデルはこれらの限界点を克服するものであり、採点者母数を含め母数の推定を安定的に行うことができる。そして分析例でも明瞭に示されたように、採点者間の個人差を識別力、困難度などの多角的な観点から評価できるという点において、その推定精度だけでなく、採点デザインを決定する上で特にその有用性があると思われる。また、採点者要因だけでなく文字の美醜など受験者に関するバイアス要因も考慮することができるモデル構成となっており、それらの効果について、希薄化の影響を抑えて推定することが可能である。

また推定法については、多くの統計モデルにおけるその利用が一般的になりつつあるMCMC法を利用し、その推定手順についての導出も行った。推定そのものは比較的簡便に行えるだけでなく母数の事前情報を加味することができ、さらに受験者数の点で問題の生じやすい小論文評価データにおいても、ベイズ法はサンプルサイズに直接的に捉われない推定法であるために、その適用は理にかなっていると言えよう。他にも、このモデルは純粋な能力評価データだけでなく、官能検査や市場調査など、測定論上の問題が生じうる評価データ全般において適用可能であり、またバイアス要因に関する共変量だけでなく潜在特性値と独立変数との回帰分析を行う文脈においても利用できるため、その広い応用可能性も期待される。

しかし、本論文で提案した項目反応モデルにも幾らかの改善点があるように思われる。特に(4)における回帰式においてはそのモデル式の関数形が仮定されているために、その仮定が真の関係と著しく乖離があった場合に、 θ_i^* 部分の評価を適切に行うことができない可能性がある。この制約は本論文で提案したモデルに限られるものではないが、例えば傾向スコア法(例えば星野・繁樹, 2004)を応用した場合などを考慮するなど、今

後の検証が必要であろう。

またこの項目反応モデルは、一般の項目反応モデルの場合と同様に、評価データの局所独立性を仮定した上で事後分布を評価している。ところが、特に一人の採点者が複数の評価観点に基づいて評価する場合、ある観点の評価がほかの観点の評価に引きずられ、その結果評価の独立性が損なわれる可能性がある。そのため、利用の際には局所独立性の問題を事前に検討しておく必要があり、その一方で局所独立性を仮定しないモデルの拡張は一つの方向性と言える。また関連して、提案したモデルの、その頑健性という観点から、既存の局所独立性を仮定しないモデル (e.g., Jannarone, 1986) を提案したモデルの形に応用した場合における、母数の推定精度の比較といった点も興味深いテーマであり、今後の検討が望まれるところである。

引用文献

- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, **22**, 47-76.
- Albert, J. H., (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, **17**, 251-269.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker Inc.
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, **12**, 86-107.
- Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, **13**, 1-18.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, **9**, 105-121.
- Chase, C. I. (1983). Essay test scores and reading difficulty. *Journal of Educational Measurement*, **20**, 293-297.
- Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, **23**, 33-41.
- Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, **57**, 373-399.
- de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov Chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, **30**, 216-232.
- Eames, K., & Loewenthal, K. (1990). Effects of handwriting and examinee expertise on assessment of essays. *Journal of Social Psychology*, **130**, 831-833.
- Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, **48**, 3-26.
- Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, **66**, 271-288.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed). New York: Chapman & Hall/CRC.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4: Proceedings of the Fourth Valencia International Meeting* (pp. 169-193). Oxford, UK: Oxford University Press.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). New York: Oxford University Press.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.
- 平井洋子 (2002). 論述的課題による高次思考能力測定の試み—採点内容の検討— *人文学報*, **326**, 17-30.
- 平井洋子 (2007). 主観的評定における評定基準, 評定者数, 課題数の効果について—一般化可能性理論による定量的研究— *人文学報*, **380**, 25-64.
- 平井洋子・渡部 洋 (1994). 小論文評点のカテゴリ

- 化に関する測定論的考察 行動計量学, **21**, 21-31. (Hirai, Y., & Watanabe, H. (1994). A psychometric study on categorization of essay examination rating. *Japanese Journal of Behaviormetrics*, **21**, 21-31.)
- 星野崇宏・繁榎算男 (2004). 傾向スコア解析法による因果効果の推定と調査データの調整について 行動計量学, **31**, 43-61. (Hoshino, T., & Shigemasu, K. (2004). Estimation of causal effect and adjustment of survey data using propensity scores. *Japanese Journal of Behaviormetrics*, **31**, 43-61.)
- Hox, J. (2002). *Multilevel analysis : Techniques and applications*. Mahwah, NJ : Erlbaum.
- Hughes, D. C., Keeling, B. F., & Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement*, **20**, 65-70.
- 伊庭幸人・種村正美・大森裕浩・和合 肇・佐藤整尚・高橋明彦 (2005). 統計科学のフロンティア 12 計算統計II—マルコフ連鎖モンテカルロ法とその周辺— 岩波書店
- Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika*, **51**, 357-373.
- Johnson, M. S., & Junker, B. W. (2003). Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding response models. *Journal of Educational and Behavioral Statistics*, **28**, 195-230.
- 梶井芳明 (2001). 児童の作文はどのように評価されるのか?—評価観点の妥当性・信頼性の検討と教員の評価観の解明— 教育心理学研究, **49**, 480-490. (Kajii, Y. (2001). How do teachers evaluate elementary school children's compositions? Validity and teachers' use of criteria. *Japanese Journal of Educational Psychology*, **49**, 480-490.)
- Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, **61A**, 273-287.
- Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago, IL : MESA Press.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, No 7.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.
- 文部科学省 (2005). 平成 18 年度国公立大学入学者選抜の概要 <http://www.mext.go.jp/b_menu/houdou/17/08/05083001.htm> (2010 年 2 月 26 日)
- Muraki, E. (1992). A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement*, **17**, 351-363.
- 大森裕浩 (2001). マルコフ連鎖モンテカルロ法の最近の展開 日本統計学会誌, **31**, 305-344.
- 大野木裕明 (1994). テストの心理学 ナカニシヤ出版
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, **24**, 146-178.
- Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability : An empirical study of a holistic rubric. *Assessing Writing*, **7**, 143-164.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : Applications and data analysis methods* (2nd ed.). Newbury Park, CA : Sage.
- Remondino, C. (1959). A factorial analysis of the evaluation of scholastic compositions in the mother tongue. *British Journal of Educational Psychology*, **30**, 242-251.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, No.17.
- 平 直樹・江上由実子 (1992). ESSAY TEST の方法論的諸問題に関する研究の動向について 教育心理学研究, **40**, 108-117. (Taira, N., & Egami, Y. (1992). Review of studies on methodological problems of essay tests. *Japanese Journal of Educational Psychology*, **40**, 108-117.
- 豊田秀樹 (2008). マルコフ連鎖モンテカルロ法 朝倉書店
- 宇佐美 慧 (2008). 小論文試験の採点における文字の美醜効果の規定因—メタ分析及び実験による検討— 日本テスト学会誌, **4**, 73-83. (Usami, S. (2008). What factor moderates the effect of handwriting quality on essay test scoring : Inves-

- tigation by meta-analysis and experiment. *Japanese Journal for Research on Testing*, **4**, 73-83.)
- 宇佐美 慧 (2009). 小論文評価データの統計解析 日本教育心理学会第 51 回総会発表論文集, 385.
- Usami, S. (2008, July). *Generalized graded unfolding model with manifest variables*. Paper presented at the meeting of the Psychometric Society, Durham, NH.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York : Springer.
- 渡部 洋 (1994). 小論文試験の特徴とその利用法について 学校教育研究所年報, **38**, 48-59.
- 渡部 洋・平 由実子・井上俊哉 (1988). 小論文評価データの解析 東京大学教育学部紀要, **28**, 143-164.
- 山内香奈 (1999). 論文評定データの解析における多相 Rasch モデルと分散分析モデルの比較 教育心理学研究, **47**, 383-392. (Yamauchi, K. (1999). Comparing many-facet Rasch model and ANOVA model : Analysis of ratings of essays. *Japanese Journal of Educational Psychology*, **47**, 383-392.)
- Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, **56**, 589-600.

付 記

本研究におけるシミュレーション及び結果部分で用いたプログラムは、統計解析ソフト R を用いて作成されたものです。利用を希望される方は宇佐美までご連絡ください。

謝 辞

本論文の作成にあたり、東京大学教育学研究科の南風原朝和先生に貴重なご指導、ご示唆を賜りました。また、株式会社ベネッセコーポレーションの鎌田恵太郎氏・矢野徹氏・島田研児氏、大学入試センターの大久保智哉先生には、小論文調査の実施の際に多くの御支援と御協力を頂きました。この場を借りて御礼申し上げます。

(2009.5.12 受稿, 11.21 受理)

A Polytomous Item Response Model That Simultaneously Considers Bias Factors of Raters and Examinees : Estimation Through a Markov Chain Monte Carlo Algorithm

SATOSHI USAMI (GRADUATE SCHOOL OF EDUCATION, UNIVERSITY OF TOKYO : JAPAN SOCIETY FOR THE PROMOTION OF SCIENCE)
JAPANESE JOURNAL OF EDUCATIONAL PSYCHOLOGY, 2010, 58, 163-175

It is generally known that evaluation of abilities through essay tests, interviews, and performance assessments may entail both rater biases, such as severity, dispersion of scores, and daily fluctuations, and examinee biases, such as expectation effects, order effects, and beauty of handwriting. In the present article, an item response model is proposed for such data, based on the Generalized Partial Credit Model (GPCM; Muraki, 1992) for polytomous responses. Effects of rater and examinee biases can be estimated directly and simultaneously through the proposed model. Parameter estimation was performed via the Markov Chain Monte Carlo (MCMC) method, which is becoming acknowledged as an effective tool for item response models. A simulation study indicated stable convergence of estimates. Additionally, actual essay test data, in which 4 raters evaluated the essays written by 304 high school students, were analyzed ; the results showed the efficacy of the proposed model.

Key Words : item response theory, evaluation of abilities, biases, essay tests, Markov Chain Monte Carlo method